Attractor networks and their relations to neural dynamics

Jake Stroud and Calvin Kao

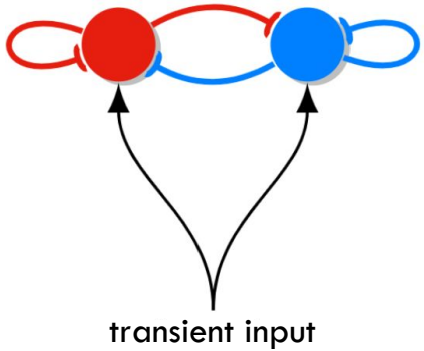
Computational and Biological Learning

UNIVERSITY OF CAMBRIDGE

What is an attractor network and why should neuroscientists care about them?

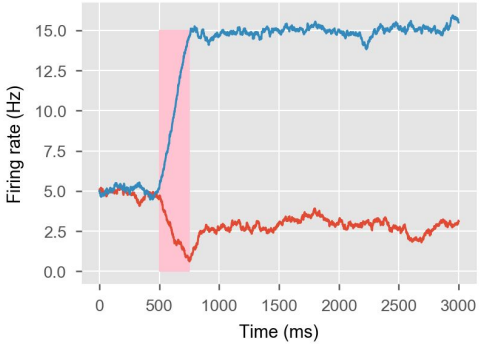
Attractor dynamics provide a mechanism for stably maintaining information over time after the removal of an input.

Attractor networks can represent either discrete (e.g., respond left vs right) or continuous inputs (e.g., angle of stimulus) in a potentially noise robust way.



Functional relevance:

- Working memory
- Associative memory
- Head direction
- Oculomotor control



Introduction to the maths behind attractor networks

Linear networks

- Stable fixed point
- Integrator

Nonlinear systems

- Discrete attractors

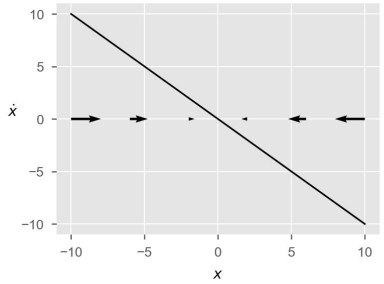
Summary and overview of models in neuroscience

Stable fixed point

A linear system is *stable* if all solutions remain bounded as $t \to \infty$.

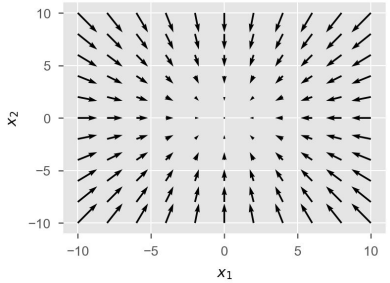
A linear system is *asymptotically stable* if all solutions converge to 0 as $t \to \infty$.

1d $\dot{x} = -x$



$x(t) = e^{-t}x(0)$

2d $\dot{\mathbf{x}} = -\mathbf{I}\mathbf{x}$



$\mathbf{x}(t) = e^{-\mathbf{I}t}\mathbf{x}(0)$

Stable fixed point

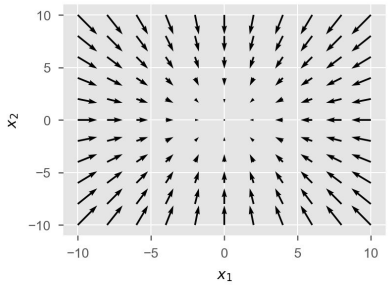
A linear system is *stable* if all solutions remain bounded as $t \to \infty$.

A linear system is *asymptotically stable* if all solutions converge to 0 as $t \to \infty$.

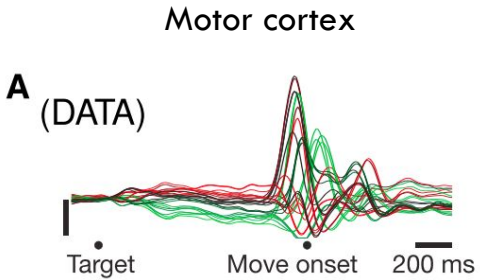
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is stable if and only if $\mathrm{Real}(\lambda_i) \leq 0$ for all i.

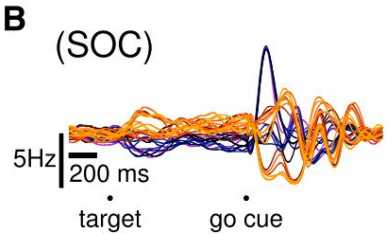
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is unstable if and only if $\mathrm{Real}(\lambda_i) > 0$ for any i.

2d $\dot{\mathbf{x}} = -\mathbf{I}\mathbf{x}$



$\mathbf{x}(t) = e^{-\mathbf{I}t}\mathbf{x}(0)$





G Hennequin et. al., Neuron, 2014.

Stable fixed point

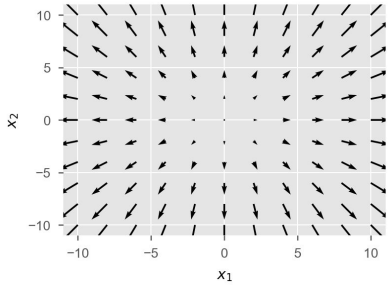
A linear system is *stable* if all solutions remain bounded as $t \to \infty$.

A linear system is *asymptotically stable* if all solutions converge to 0 as $t \to \infty$.

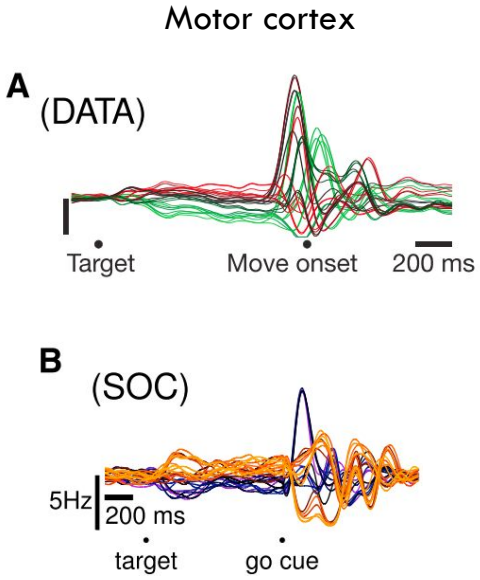
$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is stable if and only if $\mathrm{Real}(\lambda_i) \leq 0$ for all i.

$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is unstable if and only if $\mathrm{Real}(\lambda_i) > 0$ for any i.

2d $\dot{\mathbf{x}} = \mathbf{I}\mathbf{x}$



$\mathbf{x}(t) = e^{\mathbf{I}t}\mathbf{x}(0)$



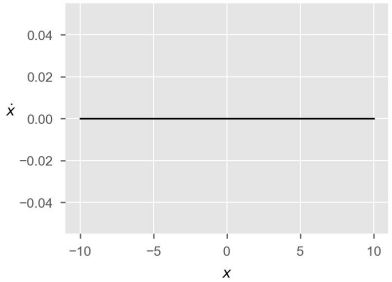
G Hennequin et. al., Neuron, 2014.

Integrator: A linear system with a line/plane attractor

How can a linear system maintain information for arbitrarily long time periods?

Set the real part of at least one eigenvalue to 0 (and all others to be less than 0).

$$\dot{x} = 0\,x$$

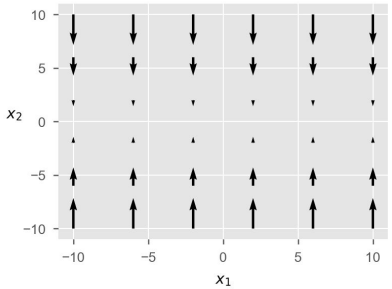


$$x(t) = x(0)$$

The integrator is a continuous attractor.

Noise is accumulated along the integrating directions.

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{x}$$

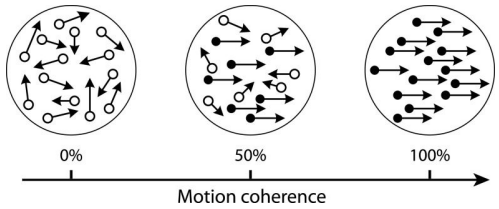


$$\mathbf{x}(t) = \begin{pmatrix} x_1(0) \\ e^{-t}x_2(0) \end{pmatrix}$$

Integrator: A linear system with a line/plane attractor

How can a linear system maintain information for arbitrarily long time periods?

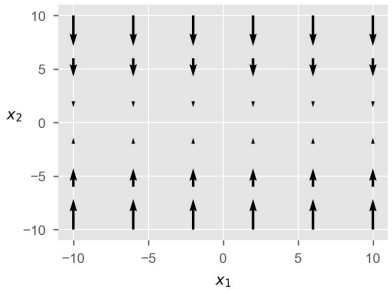
Set the real part of at least one eigenvalue to 0 (and all others to be less than 0).



The integrator is a continuous attractor.

Noise is accumulated along the integrating directions.

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{x}$$



$$\mathbf{x}(t) = \begin{pmatrix} x_1(0) \\ e^{-t} x_2(0) \end{pmatrix}$$

ME Mazurek et. al., A Role for Neural Integrators in Perceptual Decision Making, Cereb. Cortex, 2003

Summary of linear models

They can only exhibit one stable fixed point at the origin.

Integrators need a real eigenvalue at **exactly** 0.

A Proposed Neural Network for the Integrator of the Oculomotor System

Stephen C. Cannon[1], David A. Robinson[1], and Shihab Shamma[2]

Stability of working memory in continuous attractor networks under the control of short-term plasticity

Alexander Seeholzer[1], Moritz Deger[1,2], Wulfram Gerstner[1]*

Integrators don't let you forget previous inputs.

We need nonlinear systems to get multiple stable fixed points and to allow forgetting.

Linear systems are used to characterise the dynamics of nonlinear systems around fixed points.

Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks.

Sussillo D[1], Barak O.

Nonlinear system

Attractor Network Models
Wang, Encyclopedia of Neuroscience, 2009.

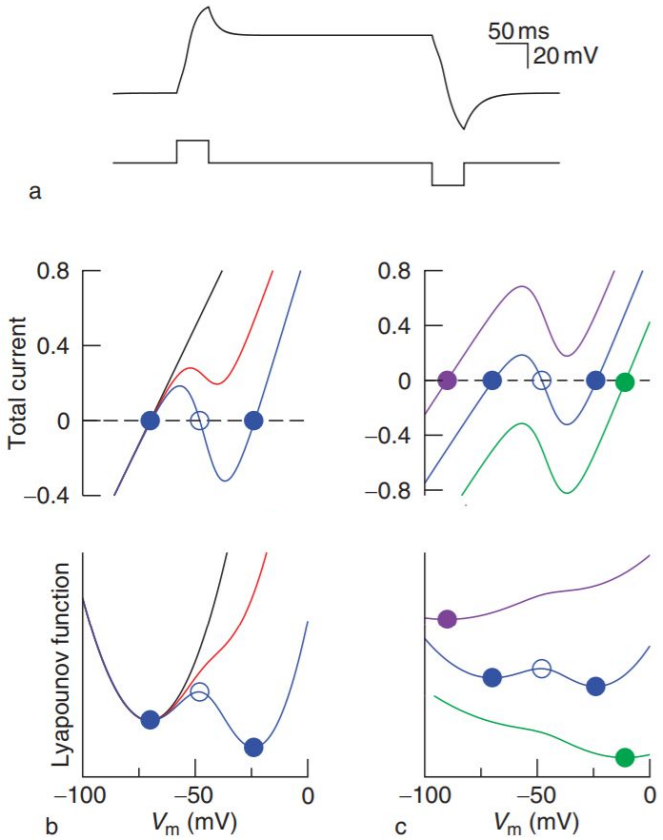
$$C_m(dV_m/dt) = -g_L(V_m - E_L) - g_{NaP}m_{NaP}(V_m)(V_m - E_{Na}) + I_{app}$$

Single neuron model as a nonlinear dynamical system with two attractor states (-70 and -25mV)

Attractor states are local minima of some “energy function”.

The attractor states depend on the input current applied, but also the parameters of the system.

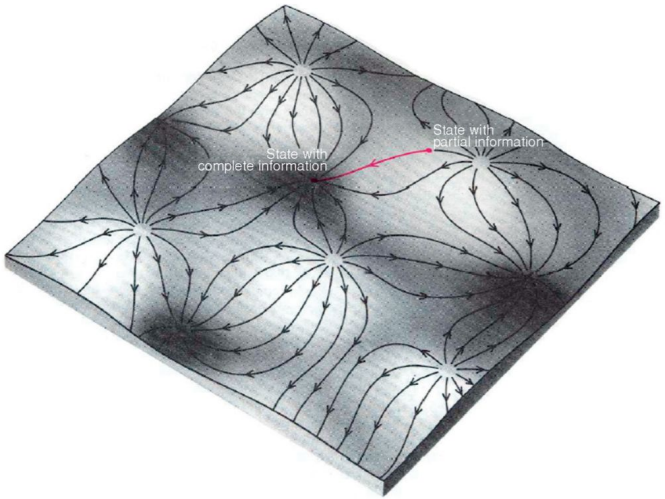
Mechanism for how neurons can “remember” past inputs for a long period of time.



Nonlinear system

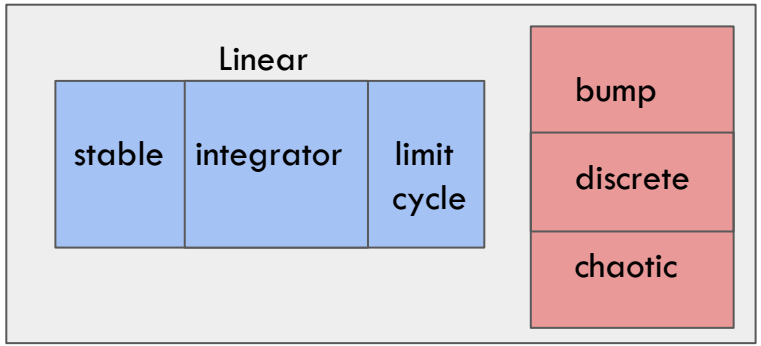
Discrete attractors as mechanisms for associative memory

Memory stored as “local minima” and recall via attractor dynamics of the network



Tank and Hopfield, 1987

Nonlinear models



Bump attractor
Linear integrator
Discrete attractor

What do we need these models for in neuroscience?

Working memory

Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory

Klaus Wimmer[1], Duane Q Nykamp[1,2], Christos Constantinidis[3] & Albert Compte[1]

Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model

Albert Compte, Nicolas Brunel, Patricia S. Goldman-Rakic, Xiao-Jing Wang

Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex.

D J Amit, N Brunel

Cerebral Cortex, Volume 7, Issue 3, April 1997, Pages 237–252,

Discrete attractor dynamics underlies persistent activity in the frontal cortex

Hidehiko K. Inagaki[2], Lorenzo Fontolan[3], Sandro Romani[3]* & Karel Svoboda[1]*

PNAS

Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex

John D. Murray[a], Alberto Bernacchia[b], Nicholas A. Roy[c], Christos Constantinidis[d], Ranulfo Romo[e,f,1], and Xiao-Jing Wang[a,b,1]

[a]Department of Psychiatry, Yale University School of Medicine, New Haven, CT 06510; [b]Department of Engineering, University of Ca... CB2 1PZ, United Kingdom; [c]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544; [d]Department of Neurobiolo...

Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity

Shaul Druckmann[1] and Dmitri B. Chklovskii[1,*]

[1]Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20176, USA

Associative memory

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. HOPFIELD

Matching storage and recall: hippocampal spike timing–dependent plasticity and phase response curves

Máté Lengyel[1], Jeehyun Kwag[2], Ole Paulsen[2] & Peter Dayan[1]

Attractor Dynamics in Networks with Learning Rules Inferred from *In Vivo* Data

Ulises Pereira[1] and Nicolas Brunel[1,2,3,4,5,*]

An attractor network in the hippocampus: Theory and neurophysiology

Edmund T. Rolls[1]

Spatial Navigation

Generation of stable heading representations in diverse visual scenes

Sung Soo Kim, Ann M. Hermundstad, Sandro Romani, L. F. Abbott & Vivek Jayaraman

Accurate Path Integration in Continuous Attractor Network Models of Grid Cells

Yoram Burak[1,2,*] and Ila R. Fiete[2,3]

Ring attractor dynamics in the *Drosophila* central brain

Sung Soo Kim[*], Hervé Rouault[*], Shaul Druckmann[†], Vivek Jayaraman[†]

See all authors and affiliations

The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep

Rishidev Chaudhuri, Berk Gerçek, Biraj Pandey, Adrien Peyrache & Ila Fiete

Oculomotor control

A Proposed Neural Network for the Integrator of the Oculomotor System

Stephen C. Cannon[1], David A. Robinson[1], and Shihab Shamma[2]

Continuous attractors and oculomotor control

H. Sebastian Seung*

Do these models actually capture neural activity?

Integrator

Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex
J Murray, A Bernacchia, N Roy, C Constantinidis, R Romo, & XJ Wang, PNAS, 2017.

Bump attractor

Bump attractor dynamics in prefrontal cortex explains behavioural precision in spatial working memory
K Wimmer, D Nykamp, C Constantinidis, & A Compte, Nature Neuroscience, 2014.

Discrete attractor

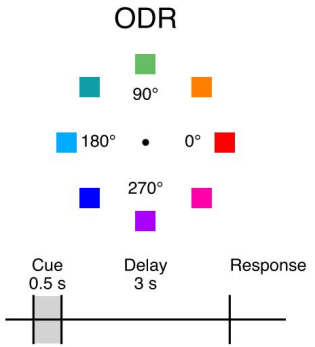
Discrete attractor dynamics underlies persistent activity in the frontal cortex
H Inagaki, L Fontolan, S Romani, & K Svoboda, Nature, 2019.

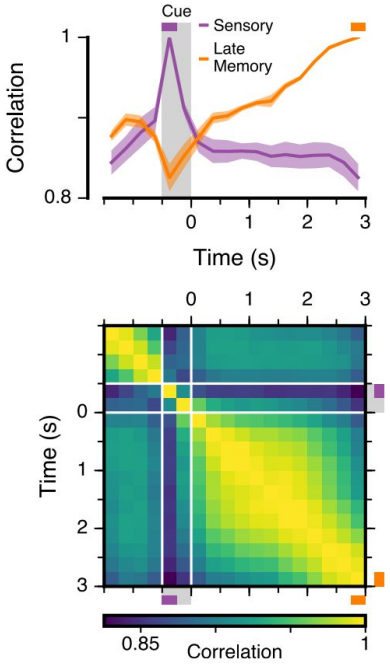
Nonlinear line attractor

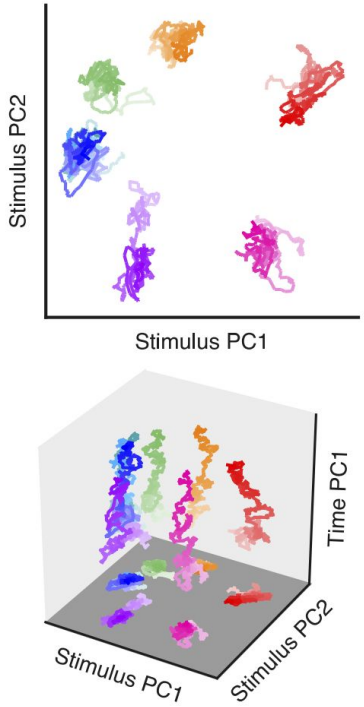
Context-dependent computation by recurrent dynamics in prefrontal cortex
V Mante, D Sussillo, K Shenoy, & W Newsome, Nature, 2013.

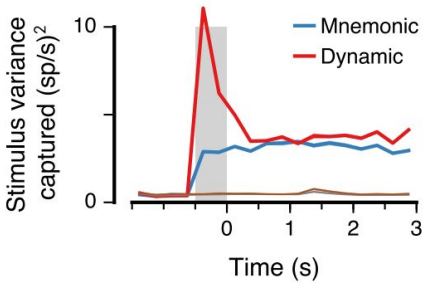
Linear integrator

Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex
J. Murray et. al., PNAS, 2017.



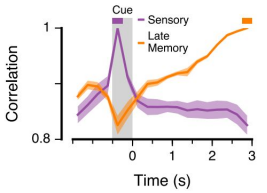


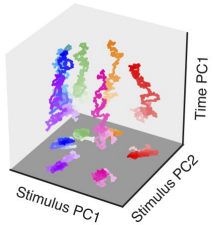


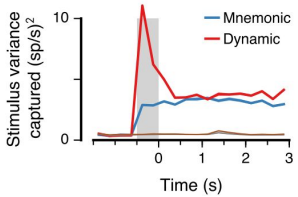


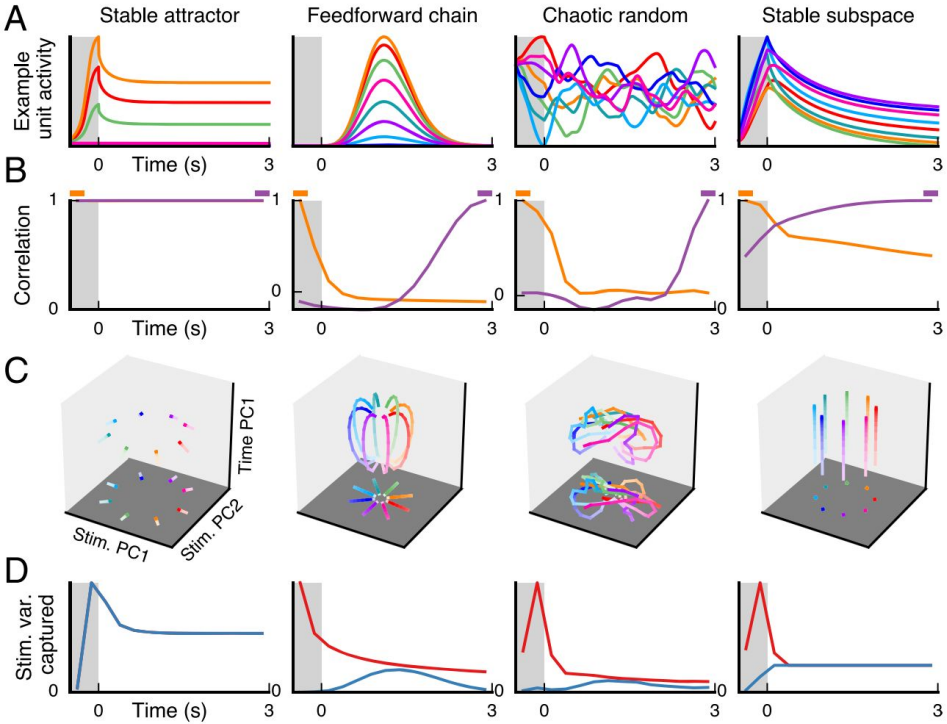
Linear integrator

Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex
J. Murray et. al., PNAS, 2017.

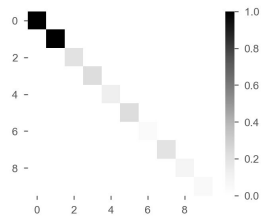








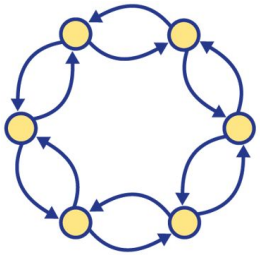
$$\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$$

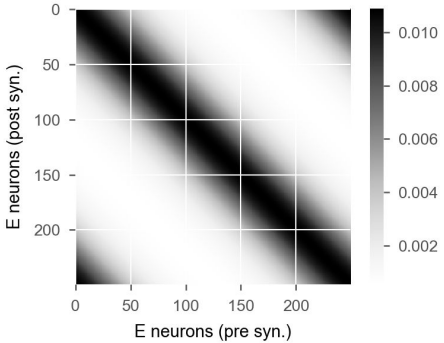


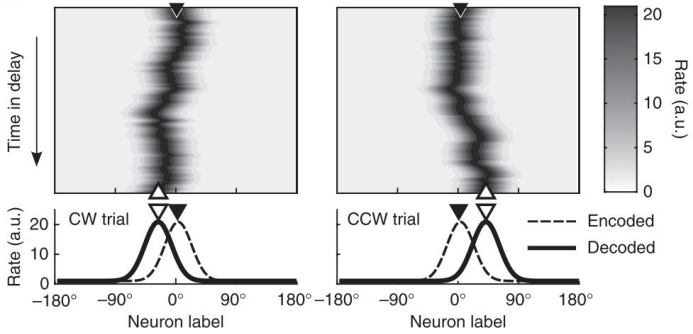
- How do you forget?
- What about the effects of noise?
- What about relating neural activity to behaviour?

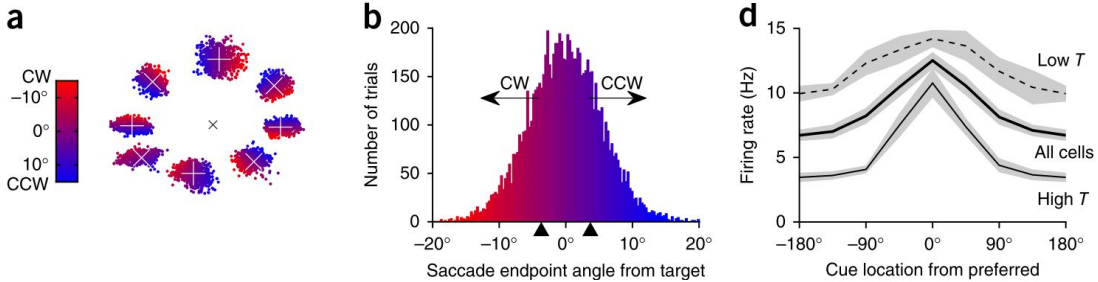
Bump/ring attractor

Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory
K. Wimmer et. al., Nature Neuroscience, 2014.







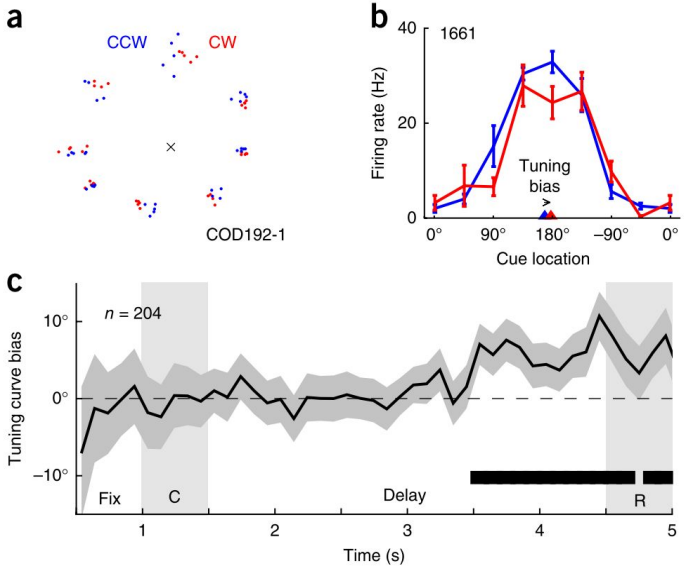


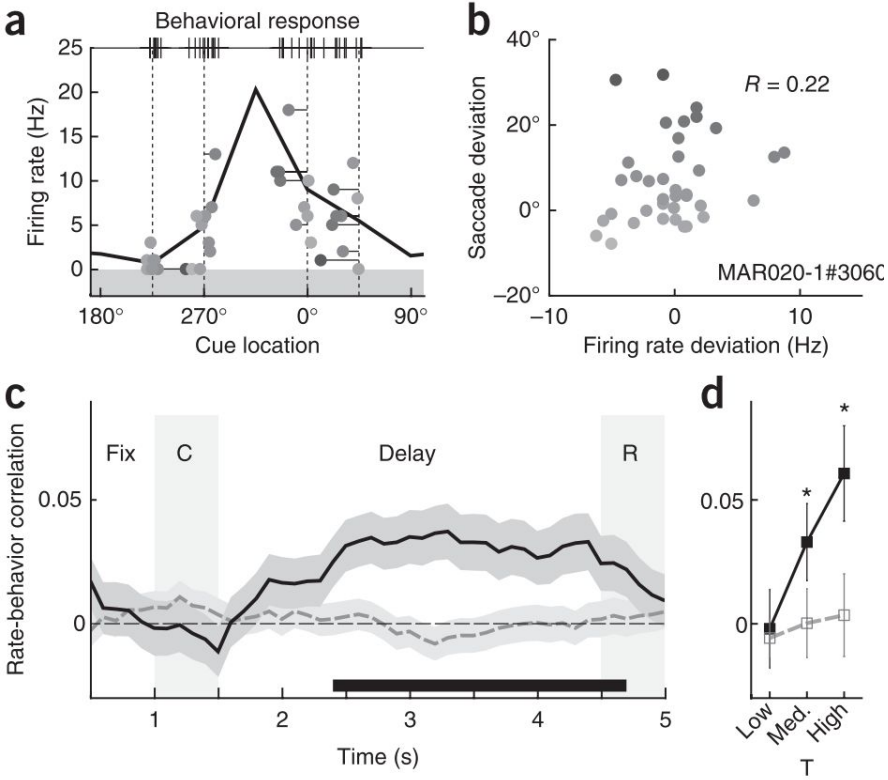
Data consists of 204 tuned neurons (out of a population of 822 neurons), 172 from monkey COD and 32 from monkey MAR.

R. Ben-Yishai et. al., Theory of orientation tuning in visual cortex, PNAS, 1995

Bump/ring attractor

Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory
K. Wimmer et. al., Nature Neuroscience, 2014.

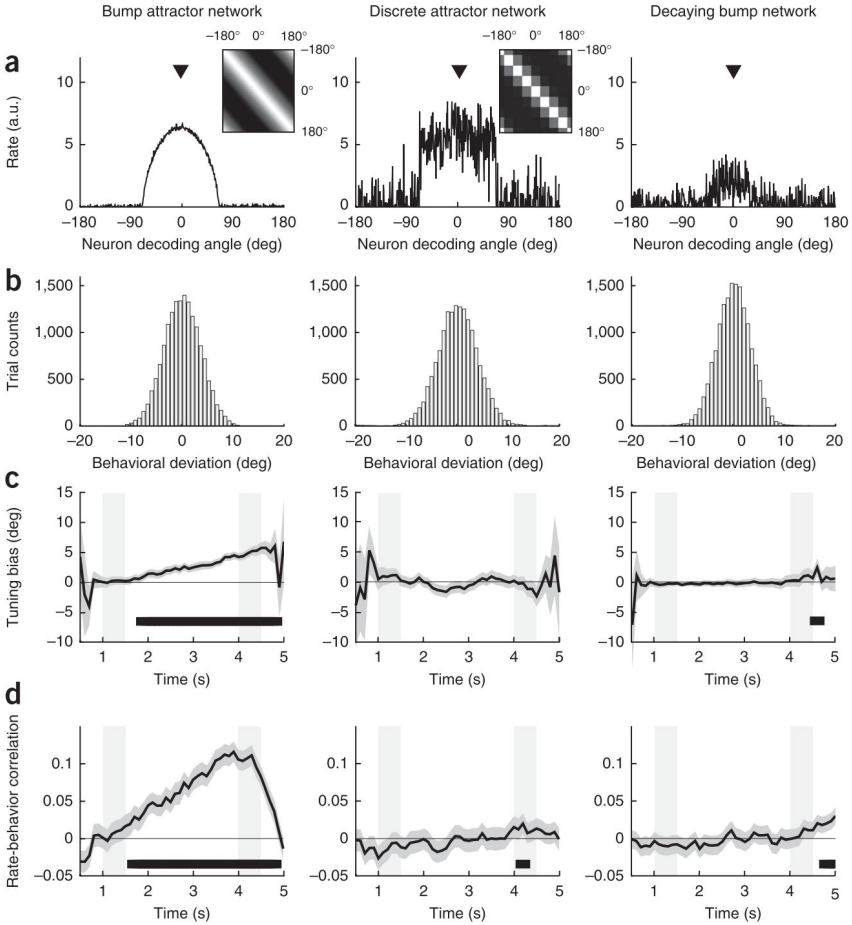




Bump/ring attractor

Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory
K. Wimmer et. al., Nature Neuroscience, 2014.

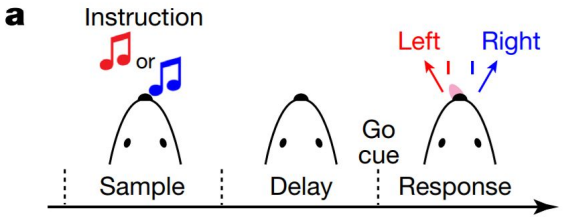
- What are the non-selective neurons doing?
- What about a population-based approach?



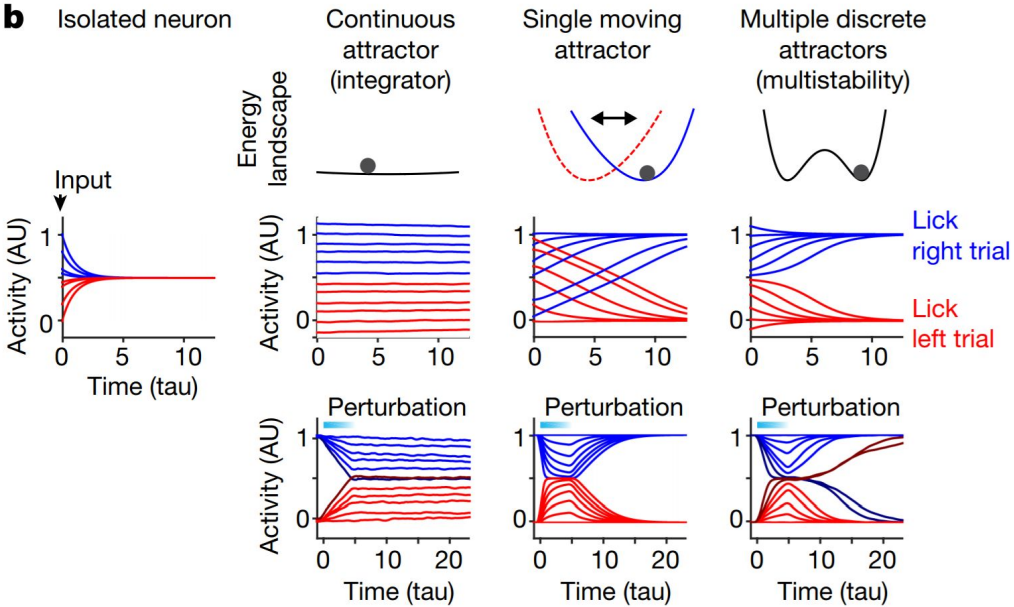
Discrete attractor

Discrete attractor dynamics underlies persistent activity in the frontal cortex
Inagaki et. al., Nature, 2019.

Tactile discrimination task with recordings from the anterior lateral motor cortex

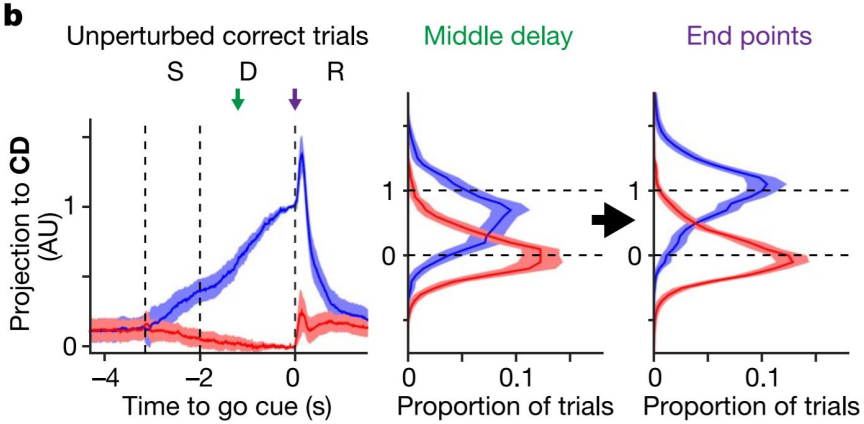


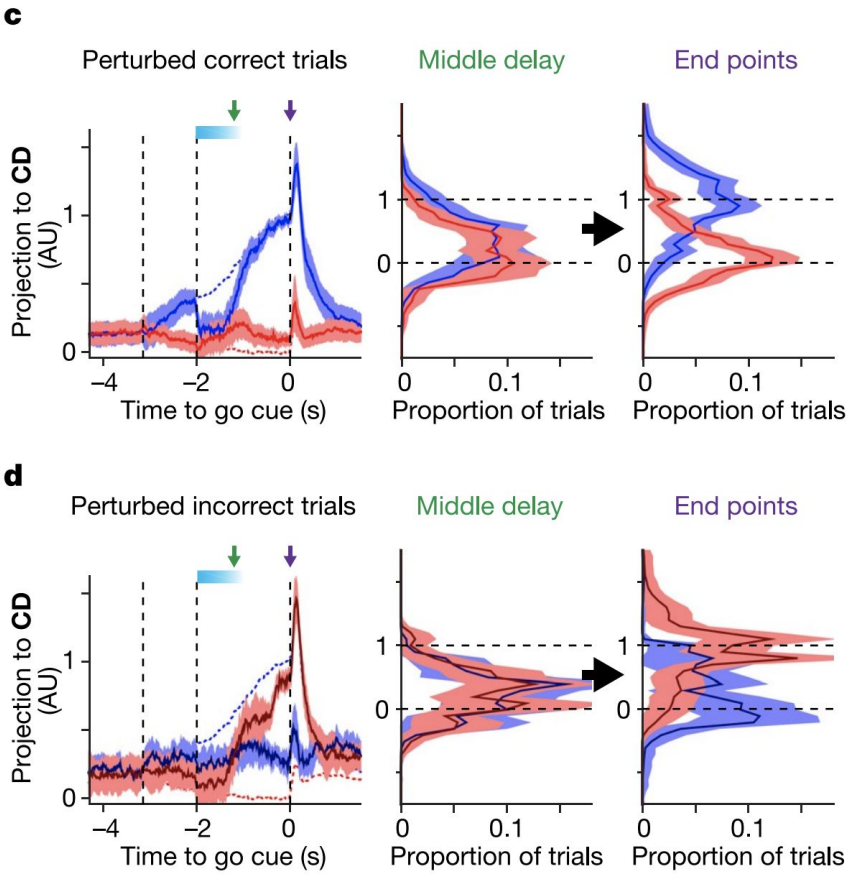
Photoinhibition experiments to distinguish between *continuous*, *single-moving*, and *multiple discrete* attractors



Discrete attractor

Discrete attractor dynamics underlies persistent activity in the frontal cortex
Inagaki et. al., Nature, 2019.





Network activity falls into the opposite attractor on incorrect trials, suggesting discrete attractor dynamics

Inagaki et al., 2019

Discrete attractor

Discrete attractor dynamics underlies persistent activity in the frontal cortex
Inagaki et. al., Nature, 2019.

Key takeaways:

1. Evidence for discrete attractor dynamics in the network
2. Optogenetic perturbations provide a powerful mechanisms for testing different attractor models

Do we need a different attractor for every movement?

	Model predictions			Data	
	Continuous attractor	Single moving attractor	Multiple discrete attractors	Fixed-delay task	Random-delay task
Funneling	+ (EDF1e)	–/+ (EDF1l)	–/+ (EDF1s)	– (Fig. 4)	N.A.
Recovery	– (EDF1c, g)	+ (EDF1j, n)	+ (EDF1q, u)	+ (Fig. 5c,g)	+ (Fig. 6f,h)
Recovery speed v.s. stim power	N.A. (EDF1f)	-/+ (EDF1m)	+ (EDF1t)	+ (EDF8c)	+ (EDF8h)
Switching	– (EDF1d, g)	– (EDF1k, n)	+ (EDF1r, u)	+ (Fig. 5d,g; EDF6i-l)	+ (Fig. 6f,h; EDF6t-x)
Phase line (slope at 1)	+ (EDF1h)	– (EDF1o)	– (EDF1v)	– (Fig. 5f)	– (Fig. 6g)

Summary so far

We have focussed on hand-crafted attractor models and how they explain neural activity.

In the last 10/15 years people have shifted to training RNNs and opening up the “black box”.

We find that trained RNNs often use attractor dynamics to solve tasks.

A diverse range of factors affect the nature of neural representations underlying short-term memory

A. Emin Orhan & Wei Ji Ma

From fixed points to chaos: Three models of delayed discrimination

Omri Barak [a], David Sussillo [b], Ranulfo Romo [c, g], Misha Tsodyks [d, a], L.F. Abbott [a, e, f]

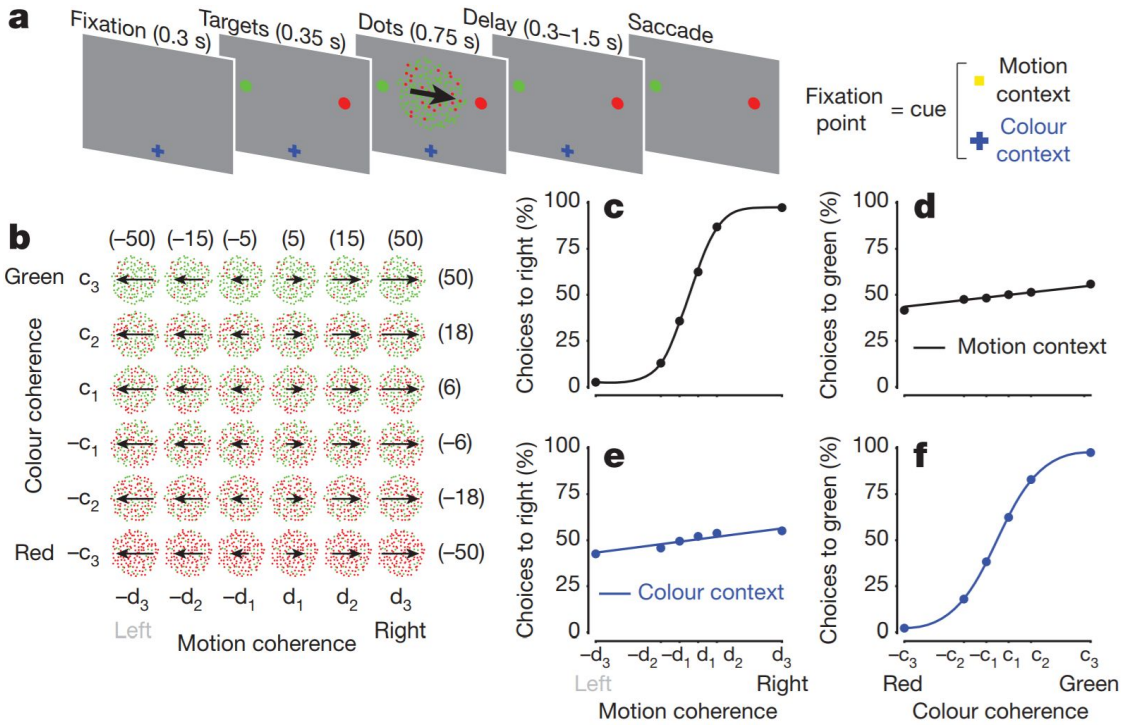
Task representations in neural networks trained to perform many cognitive tasks

Guangyu Robert Yang [1,2], Madhura R. Joglekar [1,6], H. Francis Song [1,7], William T. Newsome [3,4] and Xiao-Jing Wang [1,5*]

Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks.

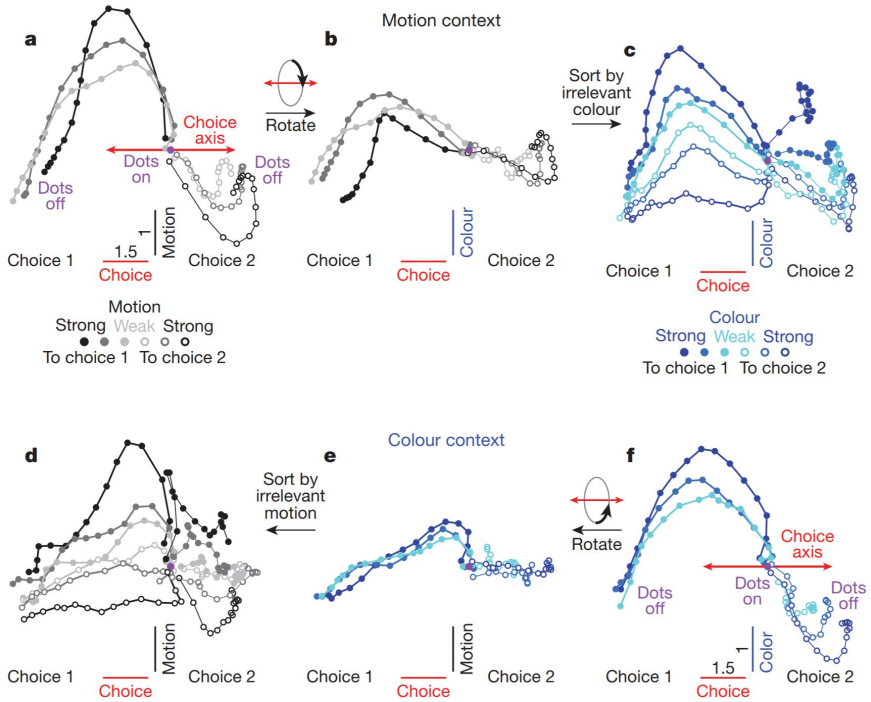
Sussillo D[1], Barak O.

Context-dependent classification (line attractors)



Context-dependent computation by recurrent dynamics in prefrontal cortex
Mante et al., Nature, 2013

Context-dependent classification (line attractors)



Movement along “choice axis” corresponds to saccade direction

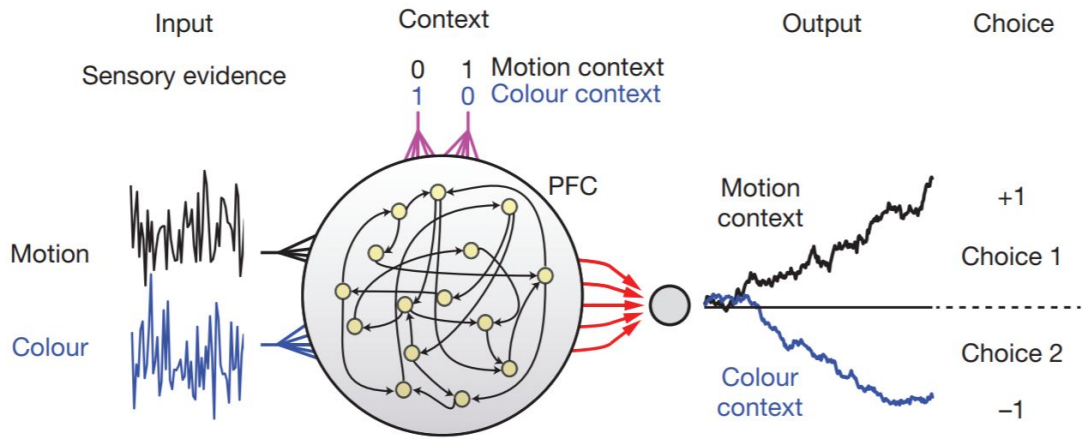
Magnitude of “arc” corresponds to strength of relevant sensory evidence

Information about the irrelevant stimuli still present, but not predictive of choice

Consistent across the two contexts

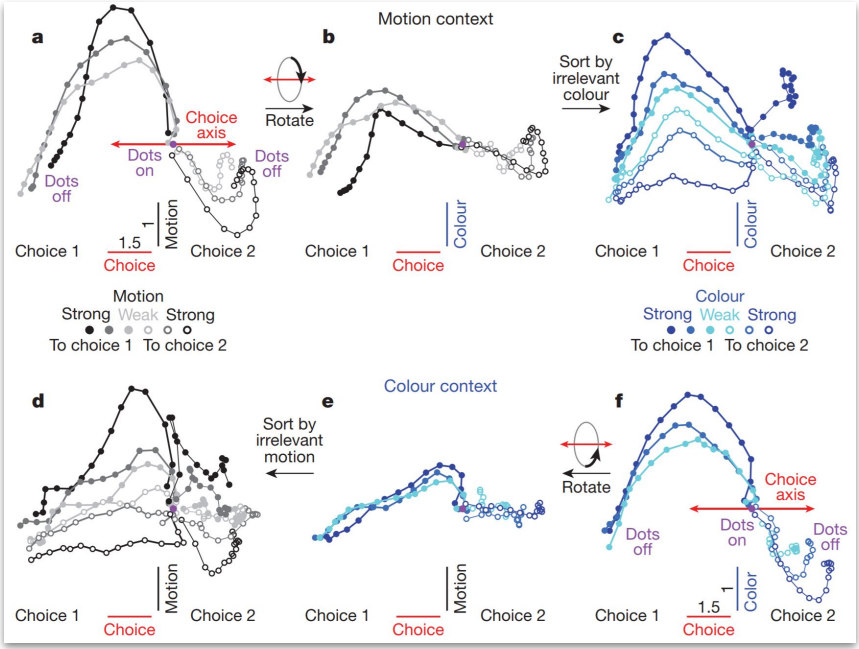
Context-dependent computation by recurrent dynamics in prefrontal cortex
Mante et al., Nature, 2013

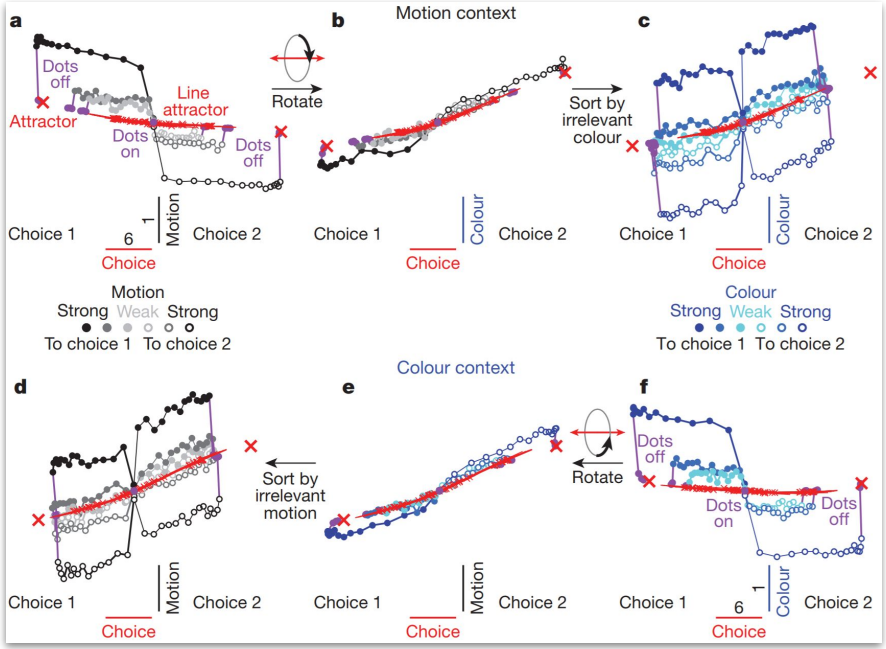
Context-dependent classification (line attractors)



Context-dependent computation by recurrent dynamics in prefrontal cortex
Mante et al., Nature, 2013

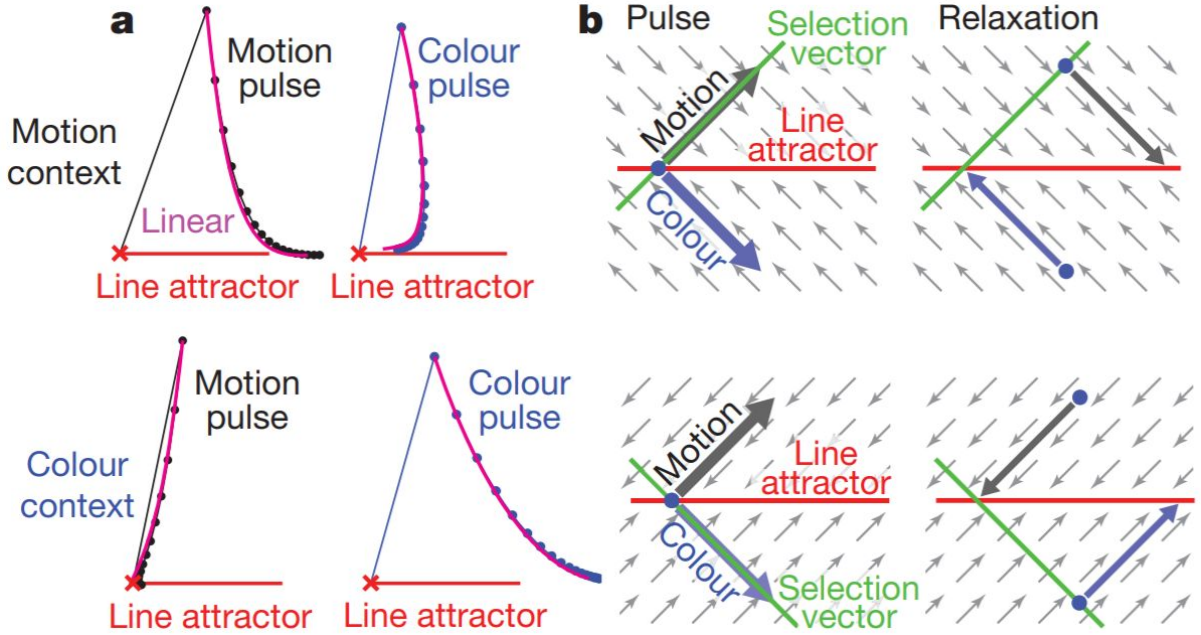
Context-dependent classification (line attractors)





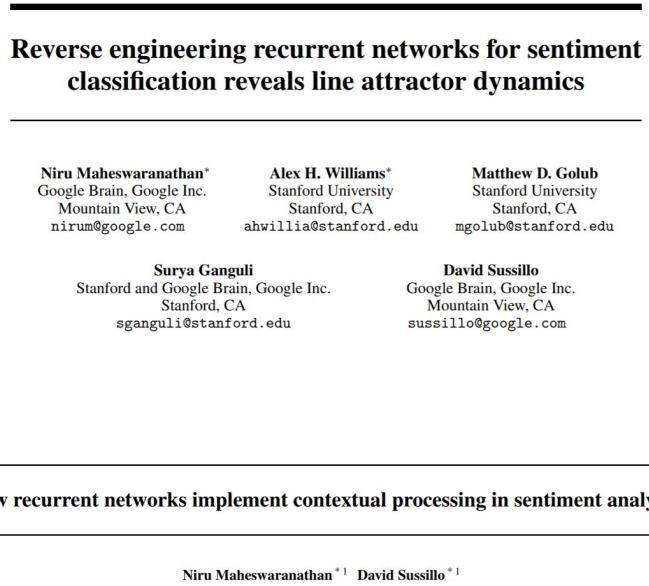
Context-dependent computation by recurrent dynamics in prefrontal cortex
Mante et al., Nature, 2013

Context-dependent classification (line attractors)



Context-dependent computation by recurrent dynamics in prefrontal cortex
Mante et al., Nature, 2013

Sentiment/context-dependent classification (line attractors)



Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics

Niru Maheswaranathan[*]
Google Brain, Google Inc.
Mountain View, CA
nirum@google.com

Alex H. Williams[*]
Stanford University
Stanford, CA
ahwillia@stanford.edu

Matthew D. Golub
Stanford University
Stanford, CA
mgolub@stanford.edu

Surya Ganguli
Stanford and Google Brain, Google Inc.
Stanford, CA
sganguli@stanford.edu

David Sussillo
Google Brain, Google Inc.
Mountain View, CA
sussillo@google.com

How recurrent networks implement contextual processing in sentiment analysis

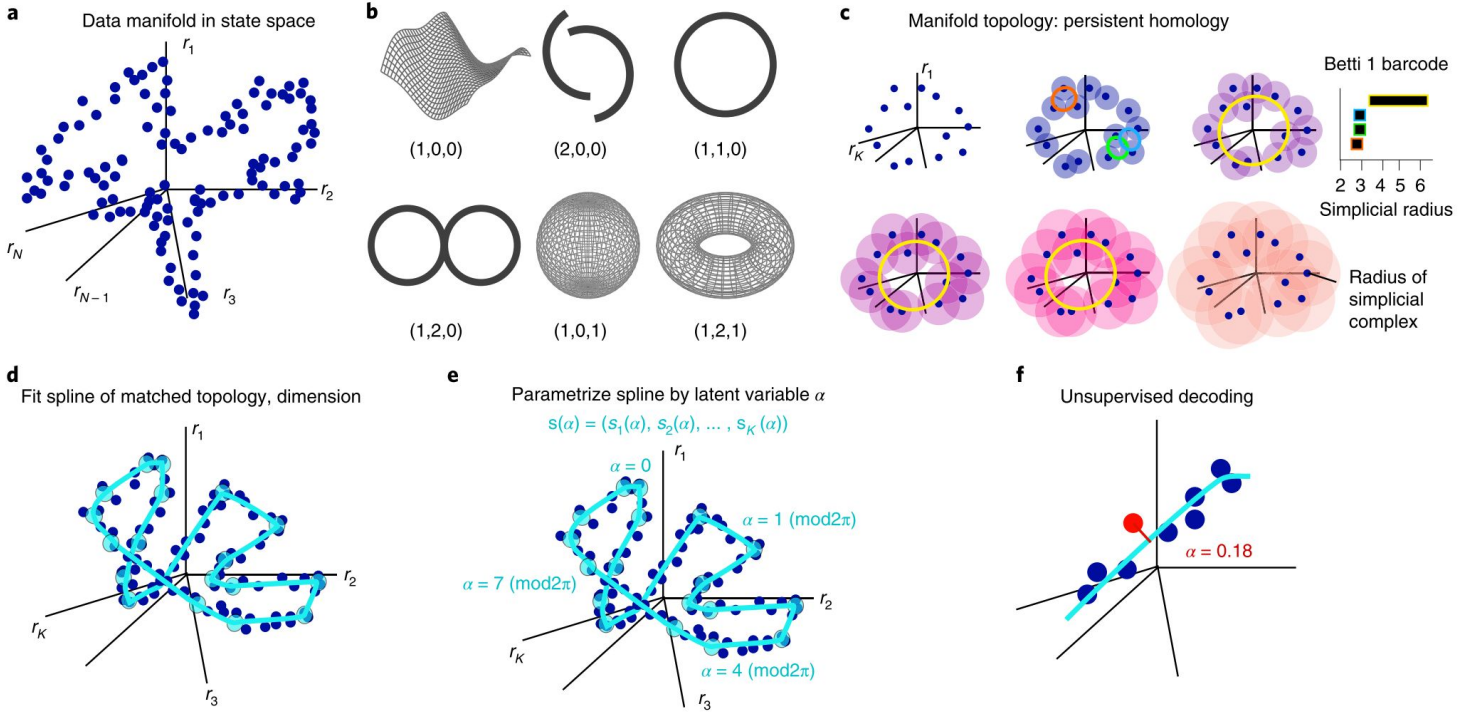
Niru Maheswaranathan [*1] **David Sussillo** [*1]

Trained RNNs to solve sentiment classification tasks (Yelp reviews)

Developed novel methods for dissecting and visualizing RNN dynamics

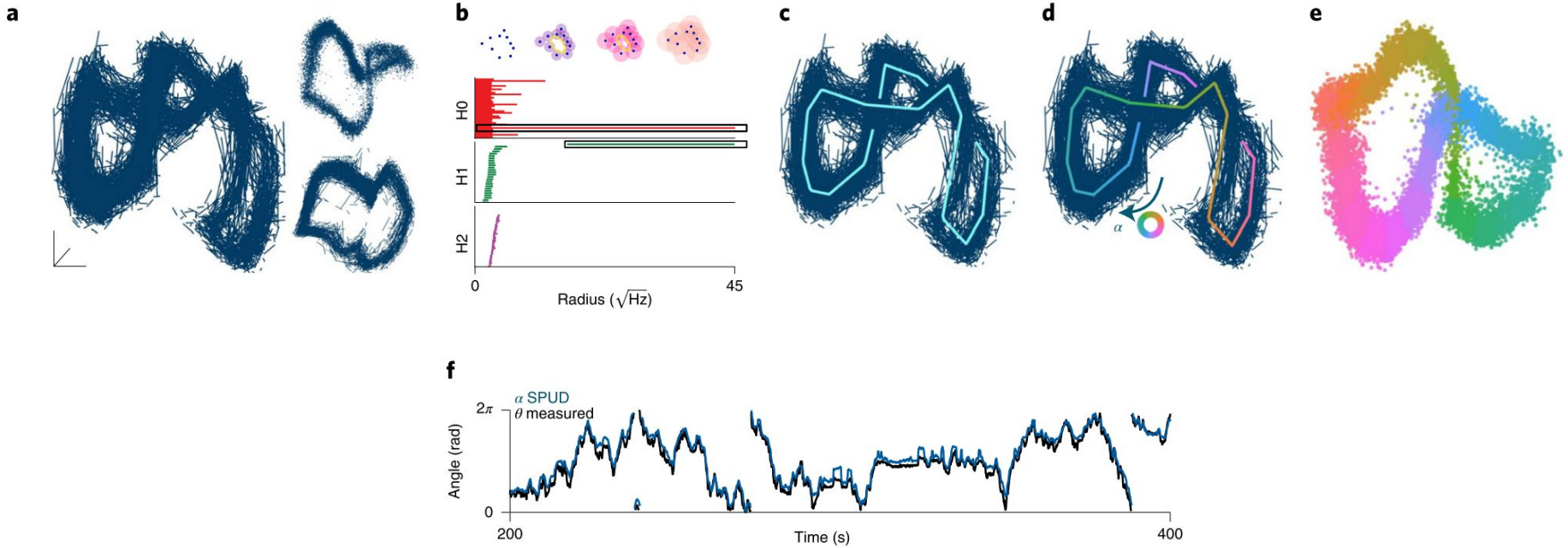
New methods for uncovering and probing attractors

The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep
R. Chaudhuri, B. Gercek, B. Pandey, A. Peyrache, & I. Fiete, Nature Neuroscience, 2019.



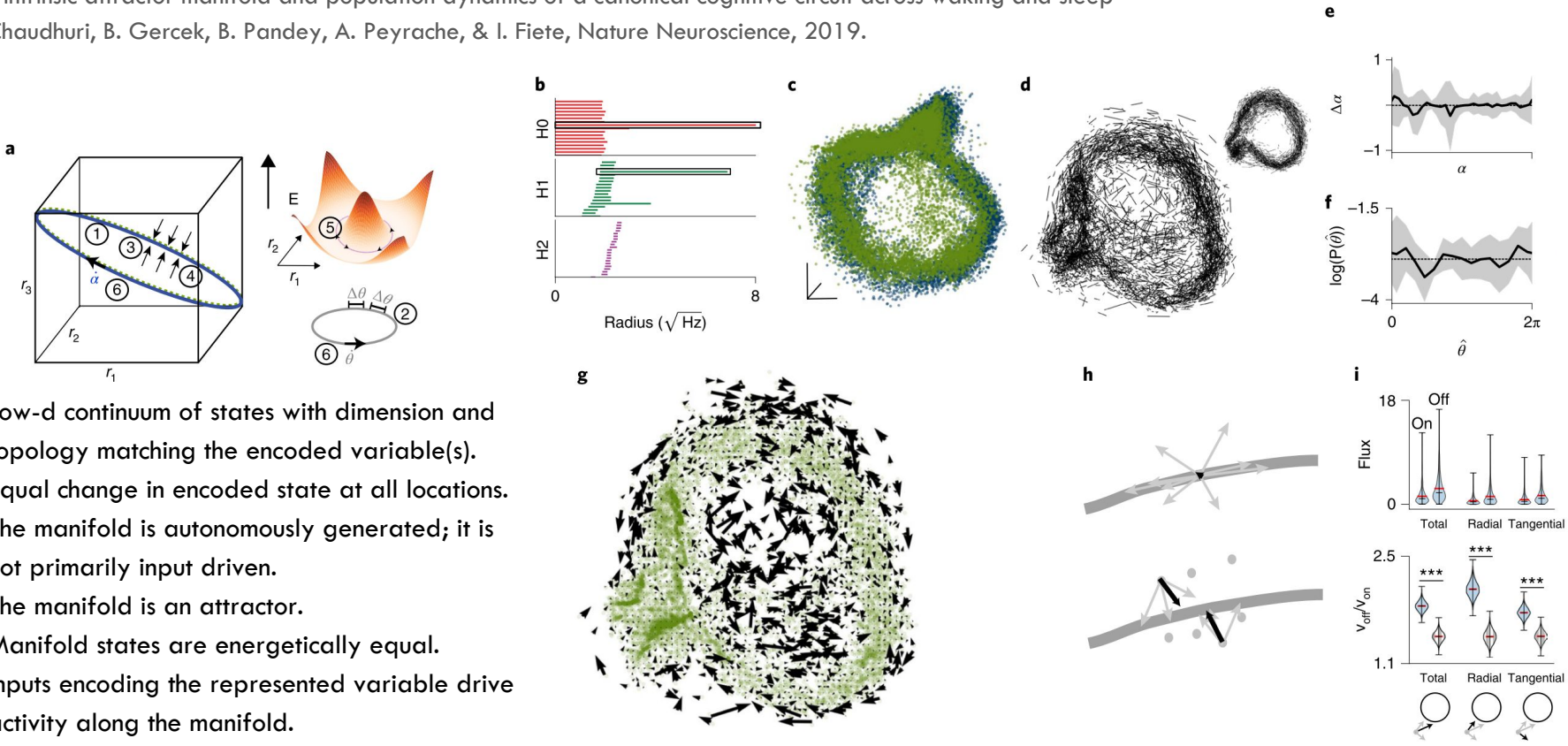
New methods for uncovering and probing attractors

The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep
R. Chaudhuri, B. Gercek, B. Pandey, A. Peyrache, & I. Fiete, Nature Neuroscience, 2019.



New methods for uncovering and probing attractors

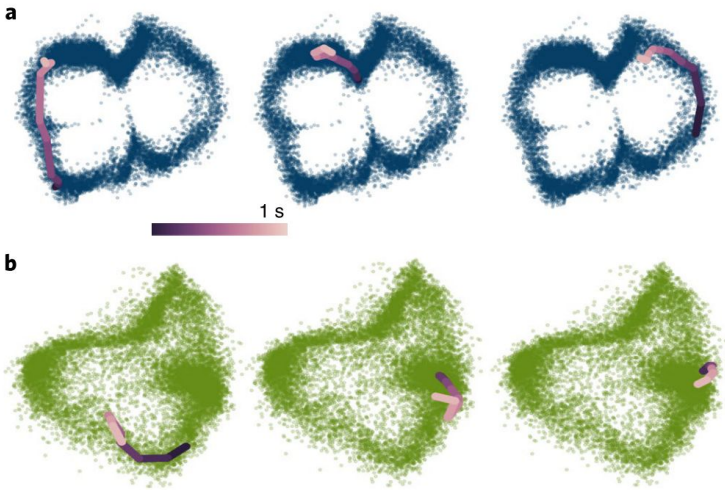
The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep
R. Chaudhuri, B. Gercek, B. Pandey, A. Peyrache, & I. Fiete, Nature Neuroscience, 2019.

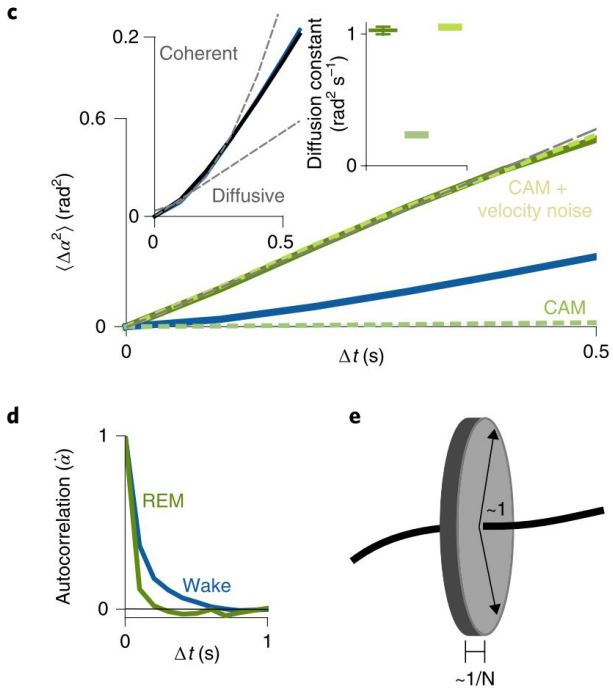


1. Low-d continuum of states with dimension and topology matching the encoded variable(s).
2. Equal change in encoded state at all locations.
3. The manifold is autonomously generated; it is not primarily input driven.
4. The manifold is an attractor.
5. Manifold states are energetically equal.
6. Inputs encoding the represented variable drive activity along the manifold.

New methods for uncovering and probing attractors

The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep
R. Chaudhuri, B. Gercek, B. Pandey, A. Peyrache, & I. Fiete, Nature Neuroscience, 2019.





Summary

- Many types of hand-crafted attractor models capture several aspects of neural dynamics from various brain regions.
- The type of attractor coincides with the type of task being performed.
- Recently, there has been a move towards population-based analyses combined with training RNNs.
- Larger scale recordings and optogenetic perturbations are proving useful for clearly elucidating network mechanisms driving neural dynamics.
- New dimensionality reduction techniques will be essential to uncover structure in neural recordings without a priori imposing structure or using task variables.